SREE

2040 Sheridan Rd
Evanston, IL 60208

Society for Research on Educational Effectiveness
Advancing Education Research

www.sree.org
847.467.4001

Abstract Title Page
Not included in page count.

Title:Power analysis for anticipated non-response in randomized block designs

Author(s): James E. Pustejovsky

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

In the ideal, a randomized control trial (RCT) provides internally valid treatment effect estimates and associated estimates of uncertainty, without relying on strong assumptions about causal mechanisms or on distributional models. This capacity for relatively model-free inference is what leads some to describe RCTs as a "gold standard" for causal inference. In practice, however, field trials on human subjects rarely meet the ideal. Missing outcome data appear commonly in social experiments, due to participant drop-out or survey non-response. In addition to reducing precision, missing data creates the possibility that treatment effect estimates may be biased (Orr, 1999). Bias arises when the process leading to missing data is non-ignorable, as when the probability that a participant drops out of an experiment positively correlated with the outcome measure of interest.

This paper examines three estimands that arise from different assumptions about the missing data process. The first assumption is that any missing outcome data are missing at random (Rubin, 1976; Little & Rubin, 2002). The predominance of analytic methods for handling missing data rely on this assumption. Missingness at random (MAR) implies that the population average treatment effect is point-identified, and is thus the appropriate estimand for inference.

The second alternative assumption is often so weak as to be nearly incontrovertible: that all outcome data are bounded by a known interval. This assumption leads to a worst-case partial identification region (PIR), a logical outer bound for a parameter (Manski, 2007). The strength of the worst-case bounds approach is that it reveals the extent to which more precise inference is based on assumptions whose credibility must be assessed by other means. Worst-case bounds have been developed for analysis of treatment effects with observational data (Manski, 1990), survey results with nonresponse (Horowitz and Manski, 1998), randomized experiments with missing data (Horowitz and Manski, 2000; Scharfstein et al., 2004, Shadish et al., 1998), and randomized experiments with non-compliance (Balke and Pearl, 1997). Tetenov (2008) studies the trade-off between statistical precision and partial identification in a simple model, very similar to that described in the paper. In addition to the width of confidence intervals, he also considers alternative loss functions and alternative decision theory models.

The third alternative assumption involves defining a less conservative PIR, using bounds for the unobservable mean outcomes relative to the observable mean outcomes. This PIR takes the form of an interval with fixed width W, centered on the observable mean outcomes. The width may be chosen directly on the basis of conservative benchmarks, or by relating it to a non-ignorable missing data model such as the pattern-mixture model described by Little (1994). This assumption is consonant with the sensitivity-analysis approach to non-randomly missing data mechanisms, as described by Allison (2001); in what follows, it is therefore referenced as the sensitivity lower bound.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

Recent guidance on the treatment of missing data in experiments advocates the use of

sensitivity analysis and worst-case bounds analysis for addressing non-ignorable missing data mechanisms; moreover, plans for the analysis of missing data should be specified prior to data collection (Puma et al., 2009). While these authors recommend only that missing data should be anticipated in the analysis plan of an experiment, it is logically consistent that anticipation of missing data should be incorporated into another important step in the planning process for an RCT: power analysis.

The paper develops power analysis formulas for the worst-case lower bound and sensitivity lower bound and compares power for these estimands to the power to detect treatment effects assuming MAR. The methods described provide a valid and principled approach for researchers to study the consequences of anticipated levels of missing data for planned experimental designs. The paper also illustrates how these methods can be used for cost-benefit analysis, weighing measurement approaches that are less expensive but yield lower response rates against approaches that are more expensive but yield higher response rates.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

Conventional power analysis formulas and programs assume that outcomes are measured on all units (Schochet, 2008; Spybrook et al., 2009). To account for missing data, researchers typically use power analysis to determine the number of complete cases necessary to achieve a given level of precision, and then inflate this number according to an anticipated level of missing data, so that the total sample size will yield adequate complete cases. This ad hoc approach fails to emphasize the assumptions necessary for its validity. The paper discusses these assumptions in detail.

Several authors have proposed methods for power analysis that use a principled approach to anticipate non-response. In particular, Hedeker et al.(1999), Jung and Ahn (2003, 2005), and Overall et al. (1998) study longitudinal designs, while Muthen and Muthen (2002) and Davey and Savla (2009) study structural equation models. However, all of these approaches assume that data are missing completely at random. I know of no literature that examines power analysis methods under alternative assumptions about the missing data process, as done in this paper.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

The paper introduces a population model that resembles the principal stratification framework of Frangakis and Rubin (2002), a randomized block sampling model, and some additional notation, all omitted here for brevity. The model is limited to an experiment that compares two treatments on a continuous outcome variable, with balanced, equally sized blocks. Matched pair designs are also excluded by assuming that each block will have at least two units in each treatment group.

I make use of three parametric assumptions in order to derive power analysis formulas: that the observed outcomes are normally distributed, with equal variance within blocks and treatment groups (A1); that the response rate in each block of each treatment group is identical (A2); and that the observed response rates are exactly equal to the population response rate (A3). Note that none of these assumptions relate to the mechanism by which outcome data are missing. Particular missingness mechanisms relate to the bivariate population distributions of the outcome and the missingness indicator, whereas A2 refers only to the marginal distribution of the

missingness indicator.

The paper establishes power analysis formulas using all three of the above assumptions, and then consider the sensitivity of the results to removal of A3 and A2 in turn. Results are presented for three different estimands, each of which is appropriate under a different assumption about the unobserved outcome data.

A typical missing data adjustment to power analysis assumes implicitly or explicitly that outcomes will be missing at random. Either the mean independence assumption discussed in Heckman et al. (1998) or the yet stronger missing at random (MAR) assumption, described by Rubin (1976), is sufficient. Either assumption point-identifies the treatment effect and implies that the observed difference in outcomes is an unbiased estimate for the average treatment effect. The paper gives a formula for power analysis of the observed outcome difference, under A1-A3. Using this formula is equivalent to conducting a conventional power analysis assuming zero non-response, then increasing the required sample size by the inverse of the response rate to account for non-response (what in the introduction I describe as the ad hoc approach). For small sample sizes in each block, it is likely that the response rate times the sample size will not be an integer, making assumption A3 less credible.

Without assuming that the observed non-response rate will be exactly equal to the population proportion of non-respondents, the above approach is only approximate. Once the data are collected, the analyst typically conditions on the observed response rates in conducting the hypothesis test. Thus, if A3 is removed, the exact power of the test can in principle be evaluated by taking its expectation over the sampling distribution of the response rates. The paper addresses several complications in this approach studies the degree to which results differs from those assuming A3. Similarly, results are presented and compared when A2 is removed also.

Next, the paper examines power analysis for a worst-case lower bound, an appropriate estimand under the extremely weak assumption that the outcome variable is bounded in a known interval (Manski, 2007). Powering a study to estimate the worst-case lower bound puts constraints on the maximum amount of missing outcome data. If the hypothesized observed treatment difference is small, the expected amount of missing data must be quite small in order for the experiment to ever establish that the true effect is bounded away from zero. Under A1-A3, power for the worst-case lower bound can be expressed in a closed form, given in the paper. If assumption A3 is removed, power depends on how the missing data process is modeled across blocks and treatment groups, in addition to how missing blocks are modeled. The paper illustrates that tests which condition on observed response rates no longer have nominal type I error rates, and gives delta-method approximations that have approximately correct type I error. The power of this estimator is examined through simulation.

Finally, the paper examines power for the sensitivity lower-bound. In certain situations, the analyst might consider a bound on the mean missing outcome, relative to the observable mean outcome. Suppose that, within each block, the mean missing outcome is within a range of W from the observed mean. This assumption partially identifies the average treatment effect, leading to a lower bound that depends on the observable treatment difference, the population non-response rate, and the width W. Just as for the worst-case lower bound, powering a study to estimate the sensitivity lower bound constrains the maximum amount of missing data. Suppose that an investigator posits an observed treatment difference of 0.25 standard deviations and a W of 0.5 standard deviations. Then an expected response rates of less than 75% implies that the lower bound is less than zero, so no amount of data will provide sufficient power to test the

hypothesis of interest. Analysis of the sensitivity lower bound largely parallels that for the worst-case bound discussed above: the paper gives power analysis formulas for this lower bound under A1-A3, as well as when A3 and A2 are removed in turn.

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

Power analysis methods that anticipate non-response are useful both to researchers designing field experiments and to granting agencies or funders who must assess the utility of proposed designs. In the design phase, these methods can be incorporated into a cost-benefit framework. The paper considers a formal model of alternative measurement designs to estimate the sensitivity lower-bound described above. It is assumed that the probability of response depends on the cost of outcome measurement: the more spent per participant, the higher the probability of response. The optimal sample size per block is derived for fixed W. A hypothetical example is given to illustrate the use of this analysis.

Granting agencies and funders can benefit from the power analysis methods described here because they lead to greater and explicit consideration of assumptions. These methods emphasize the need for reasonable (and even empirically grounded) estimates of likely response rates and levels of potential bias.

Conclusions:
Description of conclusions, recommendations, and limitations based on findings.

The appropriateness of the methods described in this paper depend on the context of the experiment being planned, and in particular on treatment under study and the method used to measure the outcome. For treatments such as dropout prevention programs that measure impacts after a considerable delay (e.g., Dynarski et al., 1998), non-random non-response may be of great concern, and should be anticipated in the planning stages using conservative assumptions. In other situations, the extreme conservatism of worst-case bounds may not be justified, but it may be prudent to use sensitivity analysis with a small value for W. Even in experiments where it is plausible that non-response is ignorably random, the assumption should be made explicit and justified during the planning stage.

Thus far I have considered only a simple randomized-block experimental design with fixed effects and normally distributed outcomes. Extensions to (block) randomized designs with binary outcomes are straightforward. When the outcome measure is binary, other estimands such as odds ratios are also relevant, as are other assumptions; for instance, Imai (2009) considers an exclusion restriction other than MAR that point-identifies the average treatment effect. The use of PIRs is less straightforward in other common experimental designs, including random effect randomized blocks or cluster-randomized designs. Application of PIRs to such designs will be explored in further work.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Allison, P. D. (2001). *Missing Data*. Thousand Oaks, CA: Sage Publications.

Balke, A. and Pearl, J.(1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association, 92 (439)*:1171-1176.

Davey, A. and Savla, J. (2009). Estimating statistical power with incomplete data. *Organizational Research Methods, 12 (2)*: 320-346.

Dynarski, M., Gleason, P., Rangarajan, A., and Wood, R. (1998). Impacts of dropout prevention programs: Final report. Technical report, Mathematica policy research, Princeton, NJ.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratication in causal inference. *Biometrics, 58 (1)*: 21-29.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica, 66(5)*: 1017-1098.

Hedeker, D., Gibbons, R. D., and Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics, 24(1)*: 70-93.

Horowitz, J. L. and Manski, C. F. (1998). Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. *Journal of Econometrics, 84 (1)*: 37-58.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association, 95 (449)*:77-84.

Imai, K. (2009). Statistical analysis of randomized experiments with non-ignorable missing binary outcomes: an application to a voting experiment. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 58 (1)*: 83-104.

Jung, S.-H. and Ahn,C. W. (2003). Sample size estimation for gee method for comparing slopes in repeated measurements data. *Statistics in Medicine, 22 (8)*: 1305-1315.

Jung, S.-H. and Ahn,C. W. (2005). Sample size for a two-group comparison of repeated binary measurements using gee. *Statistics in Medicine, 24 (17)*: 2583-2596.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika, 81 (3)*: 471-483.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data.* Hoboken, NJ: John Wiley & Sons.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review, 80 (2)*: 319-323.

Manski, C. F. (2007). *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

Muthen, L. K. and Muthen, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal, 9 (4)*: 599-620.

Orr, L. L. (1999). *Social Experiments: Evaluating Public Programs with Experimental Methods*. Thousand Oaks, CA: Sage Publications.

Overall, J. E., Shobaki, G., Shivakumar, C., and Steele, J. (1998). Adjusting sample size for anticipated dropouts in clinical trials. *Psychopharmacology Bulletin, 34 (1)*: 25-33.

Puma, M. J., Olsen, R. B., Bell, S. H., and Price, C. (2009). What to do when data are missing in group randomized controlled trials (ncee2009-0049). Technical report, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education,Washington,D.C.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63 (3)*: 581-592.

Scharfstein, D. O., Manski, C. F., and Anthony, J. C. (2004). On the construction of bounds in prospective studies with missing ordinal outcomes:Application to the good behavior game trial. *Biometrics, 60*: 154-164.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33 (1)*: 62-87.

Shadish, W. R., Hu, X., Glaser, R. R., Kownacki, R., and Wong, S. (1998). A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychological Methods, 3 (1)*: 3-22.

Spybrook, J., Raudenbush, S. W., Congdon, R., and Martinez, A. (2009). *Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software.* University of Michigan, Ann Arbor, MI.

Tetenov, A. (2009). Measuring precision of statistical inference on partially identified parameters. Working paper, Collegio Carlo Alberto.